Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models

Year: 2023 | Citations: 775 | Authors: Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu

---

Abstract

While large language models (LLMs) have demonstrated remarkable capabilities across a range of downstream tasks, a significant concern revolves around their propensity to exhibit hallucinations: LLMs occasionally generate content that diverges from the user input, contradicts previously generated context, or misaligns with established world knowledge. This phenomenon poses a substantial challenge to the reliability of LLMs in real-world scenarios. In this paper, we survey recent efforts on the detection, explanation, and mitigation of hallucination, with an emphasis on the unique challenges posed by LLMs. We present taxonomies of the LLM hallucination phenomena and evaluation benchmarks, analyze existing approaches aiming at mitigating LLM hallucination, and discuss potential directions for future research.